基于 DRN 和 FasterR-CNN 融合模型的行为识别算法 *

杨　楠，杨　莘，杜　能

(武汉科技大学 信息科学与工程学院，武汉 430081)

摘　要：针对传统单人行为识别算法易受行人形态多样性、背景和光照等影响的问题进行了研究。基于扩张卷积残差网络 DRN 在分类效果及目标检测网络 Faster R-CNN 在目标追踪方面的准确性，提出了一种 DRN 和 Faster R-CNN 的融合网络模型。该模型在 Faster R-CNN 中融入 DRN 的扩张卷积残差块代替原来的一般卷积层部分。并对融合模型进行了两方面的改进：在每一层前面添加一个 batch normalization 层；用三层扩张卷积残差块代替部分两层残差块。实验结果表明三种融合网络识别算法在 Olympic sports dataset 数据库上较其他行为识别算法取得了更高的 mAP。其中，包含三层扩张卷积残差块的融合模型识别性能最好，mAP 达到 78.9%。

关键词：行为识别；DRN；Faster R-CNN

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2018.05.0354

Behavior recognition algorithm based on DRN and Faster R-CNN fusion model

Yang Nan, Yang Shen, Du Neng

(*School of Information Science & Engineering*, *Wuhan University of Science &Technology*, *Wuhan* 430081, *China*)

Abstract: Due to the traditional single person behavior recognition algorithm is easily affected by the diversity, background and illumination of pedestrians. Based on the accuracy of convolution residual network DRN in classification and detection network Faster R-CNN in target tracking, we proposes a fusion network model composed of DRN an Faster R-CNN. The model is integrated with dilated convolution residual in Faster R-CNN to replace the original convolution layer. We also made two improvements to the fusion model, add a Batch Normalization layer in front of each layer; Used three levels of dilated convolution residual blocks instead of partial two levels of residual blocks. The experimental results show that the three fusion network recognition algorithms proposed in this paper have achieved a higher mAP than other behavior recognition algorithms on the Olympic Sports Dataset database. Among them, the fusion model with three layers of convolution residual blocks has the best recognition performance, and mAP achieves 78.9%.

Key words: Behavior recognition; DRN; Faster R-CNN

0　引言

近些年来，人体行为识别在智能视频监控、视频检索和人机交互等多种应用中引起了广泛的关注[1]。目前国内外对人体行为识别都投入了大量研究，也取得了一定进展，但复杂的背景、照明变化、外观差异和运动行为繁杂等因素使得人体行为识别成为具有挑战性的任务，所以目前行为识别的准确度并不能满足实用化的需求。学者们提出了多种行为识别算法，其中基于机器学习的方法吸引了广泛的关注[2]。

针对行为识别的研究，传统的机器学习算法一般由特征提取和行为分类两个部分组成，常见特征提取算法有 LBP(local binary patterns)、HOG(histogram of oriented gradients)[3] 和 SIFT(scale-invariant feature transform)[4]等。常见的分类器则有决策树、支持向量机(support vector machine,SVM) [5,6]等。传统的行为识别算法采用提取特征再利用分类器进行分类，存在提取特征不全面，人力消耗过大等问题。

近几年，卷积神经网络已经广泛应用于图像分类和目标识别等任务中。在 ImageNet ILSVRC 中的图像分类比赛中，2012 年由 Krizhevsky 等人[7]实现的 AlexNet 卷积神经网络以 16%的错误率夺得比赛的冠军，并使得卷积神经网络在计算机视觉领域受到广泛的关注。在之后的比赛中，各类卷积神经网络层出不穷，由 He 等人[8]实现的残差网络(ResNet)则是 ILSVRC2015 的冠军模型，ResNet 的跳跃式链接能有效解决较深网络中“退化”的问题。扩张残差网络(dilated residual networks, DRN) [9]则是在 ResNet 的基础上结合了扩张卷积(dilated convolutions)的算法。该算法通过增大卷积的感受野(receptive field)从而达到替代

收稿日期：2018-05-06；**修回日期：**2018-06-26　　**基金项目：**国家自然科学基金资助项目（61502358）

作者简介：杨楠（1993-），男，湖北荆州人，硕士研究生，主要研究方向为图像处理、机器学习（330903312@qq.com）；杨莘（1977-），女，湖南涟源人，副教授，博士，主要研究方向为多媒体信号与信息处理；杜能（1997），女，湖北黄冈人，本科，主要研究方向为图像处理、机器学习．

池化层的目的，在维持原网络的感受野不变的同时又不会损失图像空间的分辨率，从而能够最大限度的保留输入图像中的细节信息。目标识别任务则可以划分为目标的追踪和行为的分类两部分，传统基于机器学习的行为识别算法一般没有进行目标的追踪，而是直接对行为进行分类，如依据时间、空间建模[10,11]直接对待测行为进行识别；人为提取行为特征再调用分类器进行分类[12]；文献[13]采用 HOG+CNN 进行特征的提取，并通过时间排序结合支持向量机(TOI+SVM)进行行为分类。目前，将目标的追踪和分类同时进行的算法在各种数据库中取得了前所未有的成果[14~17]。由 Girshick 等人[15]提出的基于 R-CNN(regions with CNN)目标检测算法将目标检测的平均分类精度由 34.3%提升到 66%。但 R-CNN 训练步骤较为复杂，且测试时间较长。针对这些问题，研究人员相继提出了 Fast R-CNN[16]，Faster R-CNN[17]等算法，这两种算法采用卷积神经网络来进行目标的追踪和分类，不仅解决了 R-CNN 算法的检测耗时较长的问题，而且有效提升了目标检测的平均分类精度(mean average precision, mAP)。本文结合 DRN 在分类任务上的准确性以及 Faster R-CNN 在目标追踪上的精确度，融合成一个新的网络模型完成行为识别任务。

1 深度卷积神经网络模型

CNN 在普通神经网络的基础上，添加了能够实现卷积操作的卷积层和进行降采样的池化层。在卷积层中，每一个神经元只与上一层的部分神经元相连。每一个卷积层通常包含多个滤波器，即特征平面，每个滤波器包含 n*n 个神经元，n 为大于等于 1 的数，对于上一层输入网络，经过每个滤波器的神经元共享权值，该权值即为卷积核。下面简要介绍 DRN 网络和 Faster R-CNN 网络的模型结构。

1.1 DRN 网络

DRN 是残差网络(ResNet)的一种变体。ResNet 是由何恺明等人实现的一种特殊的残差网络，即跳跃链接型网络。随着神经网络的深度不断加深，模型的学习能力会在某个深度达到稳定，继续增加模型的层数时，模型前面一个细小的改变都会在模型后面引起很大的变化，即会出现“梯度消失”或“梯度爆炸”现象，此外，还会产生 “退化”问题，即网络层数很深时，其学习能力不仅不会提升反而下降，此时训练准确率和测试准确率均在下降，深度网络就变得难以训练了，且这种学习能力的下降与过拟合无关。针对所谓的“退化”问题，DRN 提出一种 Residual 结构，如图 1 所示。图中给出了一个两层的残差学习模块，即一个 Residual 的结构中含有两层卷积层，其中 x 为输入，relu 为线性整流函数（Rectified Linear Unit, Relu），也称为修正线性单元，是机器学习中较为普遍的激活函数（activation function）[18]，映射函数 H(x)由 F(x)改为 F(x)+x。

DRN 在 ResNet 的基础上加入了扩张卷积的思想，通过扩张卷积可以在实现与原网络中一致感受野的同时保持输出尺寸与输入一致，且无须经过池化操作。实现过程如图 2 所示，(a)表示扩张为 1 的卷积，与普通卷积操作无异，(b)(c)分别表示扩张为 2 和 4 的卷积操作。由于扩张卷积可以代替池化层，该算法在增大卷积感受野的同时能保持输出与输入尺寸一致，从而能够最大限度的保留输入图片中的细节信息，使 DRN 较 ResNet 在图片分类上的性能有了一定的提升。

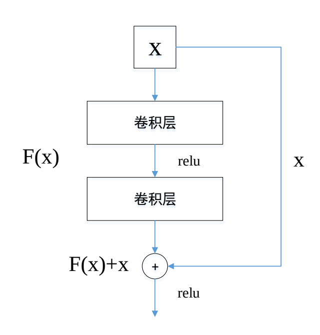


图 1 残差块结构

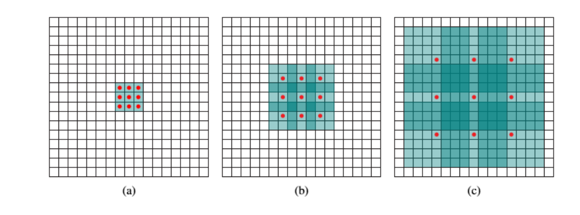


图 2 扩张卷积实现过程

1.2 Faster R-CNN

针对 R-CNN 和 Fast R-CNN 中 selective search 算法生成目标建议框的速度问题，Faster R-CNN 引入了区域建议网络(region proposal network, RPN)代替 Selective Search 算法用于生成目标建议框[17]，极大地提升了目标建议框的生成速度。该部分的网络结构如图 3 所示。

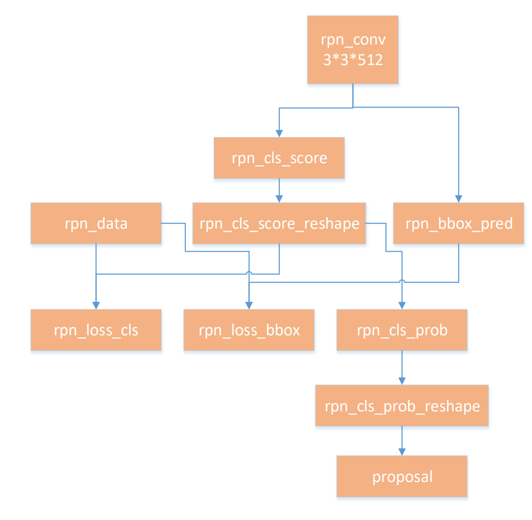


图 3 RPN 网络结构

RPN 网络进行第一个卷积操作之前，在输入的每个点上都形成三种尺寸、三种比例的 anchor，每个 anchor 在原图对应 9 个目标框。然后，原图得到的 9 个目标框在图片上以步长为 16 扫描全图，每步都得到 9 个目标框，扫描结束后得到的全部目标框数量一般在 2 万~4 万。剔除跨越边界的目标框，剩下的 6000~10000 个目标框作为目标建议框带入 RPN 网络中进行训

练。然后，将输入特征带入 RPN 网络开始运算，第一个卷积层采用 512 个 3*3 的滤波器，步长为 1，填充为‘SAME’，此时能保持输入和输出尺寸一致，激励函数为 Relu。“rpn_cls_score”和“rpn_bbox_pred”为两个全连接层，分别输出目标框在前景目标上的得分和在回归信息。两个包含“reshape”的层为维度转换层，能根据需要将输入的维度进行变换。“rpn_cls_prob”层为一个 softmax 层。“proposal”为目标框生成层，该层中剔除跨越边界的目标框，并通过非极大值抑制(non-maximum suppression, NMS)[19]结合目标框前景得分筛选部分目标框，最后通过目标框的回归信息得到 RPN 网络给出的目标建议框，最后选取 256 个目标建议框作为 RPN 网络的输出。

“rpn_loss_cls”和“rpn_loss_bbox”分别对应于 RPN 网络的得分损失值和回归损失值。将得分损失和回归损失按一定的权重相加即为 RPN 网络的损失，其损失函数的定义为：

$$L(\{p_i\},\{t_i\})=\frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \qquad (1)$$

其中：i 为 anchor 的索引，λ 为 10，p_i 表示网络对索引为 i 的 anchor 对应目标框预测为目标的概率值。p_i^* 是正确标注(Ground Truth,GT)目标框的概率，只能为 0 或 1，若该目标框为前景目标，p_i^* 为 1；若该目标框为背景，p_i^* 为 0。t_i 为一个向量，表示预测目标框左上角和右下角四个坐标值，t_i^* 为 GT 目标框左上角和右下角的四个坐标值。式(1)中，分类损失 $L_{cls}(p_i,p_i^*)$ 是目标和非目标的对数损失，其损失函数公式为

$$L_{cls}(p_i,p_i^*)=-log[p_i^* p_i+(1-p_i^*)(1-p_i)] \qquad (2)$$

回归损失 $L_{reg}(t_i,t_i^*)$ 表达式如下：

$$L_{reg}(t_i,t_i^*)=R(t_i-t_i^*) \qquad (3)$$

其中：$R(x)$ 是文献[13]中定义的鲁棒损失函数，其表达式如下：

$$smooth_{L1}(x)=\begin{cases}0.5x^2 & if\ |x|<1\\ |x|-0.5 & otherwise\end{cases} \qquad (4)$$

2　基于 DRN 与 Faster R-CNN 融合模型的行为识别算法

2.1　数据预处理

本文采用 Olympic sports dataset 数据库，该数据库于 2010 年由斯坦福大学发布，分为篮球、举重、长跑等十六类体育运动，其中每一类有 50 个视频，部分示例见图 4。本文在彩色数据库中挑选 5000 幅图像并进行 GT 目标框标注，同时将这些图像左右翻转得到 5 000 张同样含有 GT 目标框的镜像图像。将这 10 000 幅图像作为训练集。从原彩色数据库中另外挑选 2 000 幅图像作为交叉验证集，5000 幅作为测试集。在通过交叉验证集调整超参数及测试集检测训练结果时无须对目标框的位置进行验证，只需检测行为识别的准确性，因此交叉验证集和测试集无须手动标注 GT 目标框。

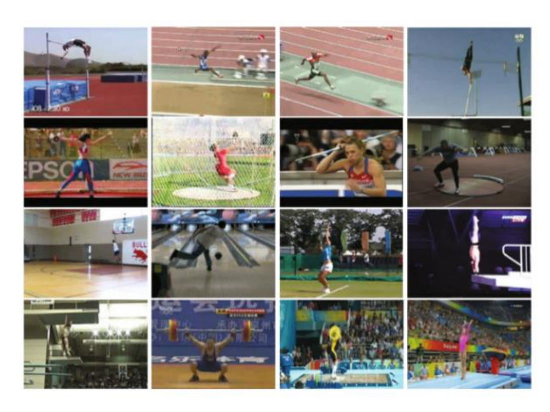


图 4　Olympic sports dataset 数据库示例

2.2　融合网络结构

图 5 给出了融合的网络模型，该融合模型中橘黄色部分与 1.2 节 Faster R-CNN 模型中的 RPN 网络一致。金色框部分的“Roi_data”层存储了 RPN 网络对输入图片推荐的感兴趣区域(Region of Interest, Roi)。“Roi_pool5”为 Faster R-CNN 网络中提到的感兴趣区域(Region of Interest, Roi)池化层，其主要作用是将全连接层的输入尺寸调整一致，都为 7*7。

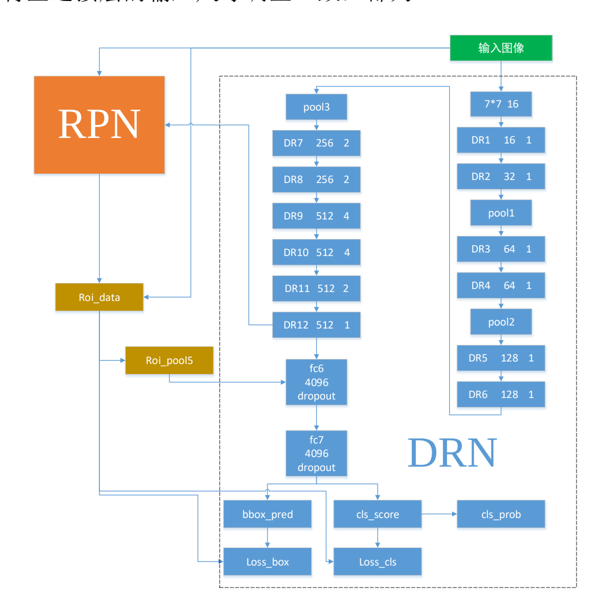


图 5　融合网络结构

融合模型中的虚线框为融合模型的主体，即 Fast R-CNN 部分，本文采用 DRN 网络来代替原 Faster R-CNN 中所用的 VGG16 网络[20]。由于 Roi 池化层可以将所有通过该层的尺寸变为一致，所以该网络的输入图像不唯一。蓝色部分第一层为 7×7 的卷积层，填充(padding)为 2，步长(stride)为 2，16 代表用于该层中卷积的滤波器个数，由于步长为 2，输入经过该卷积层后尺寸变为原来的一半，融合进来的 DRN 网络，除去第一层为一个单独的卷积层和前半部分的三个池化层外，其余的部分皆为 2 层的扩张卷积残差块，其结构如图 1 所示，所有的扩张卷积残差块中两层卷积层皆为 3*3 的滤波器，步长都为 1，填充皆为“SAME”方式。卷积层的滤波器个数以及卷积层中的扩张值不一致。图 5 中蓝色部分第二层中“DR1”代表该模型中的第一个扩张卷积残差块，每个扩张卷积残差块皆由两个扩张卷积层组成，数字 16 为滤波器数量，两个卷积层的滤波器个

数都为 16，1 表示扩张值，两个卷积层的扩张值为 1，即图 2 所示，采用 1-dilated 的处理方式。下面各层的扩张卷积残差块中数字含义相同。该模型中“pool1”，“pool2”，“pool3”是三个相同的最大值池化层，池化层尺寸都为 2*2，步长为 2，填充采用的方式为“SAME”，经过池化层后的输出尺寸为输入的一半。在第三个池化层之后，即从第 7 个扩张卷积残差块开始采用扩张为 2 或者 4 的卷积，如图 2(b)(c)中所示，扩张卷积增大了感受野，代替了池化层，避免了尺寸的不断减小，因此第 3 个池化层后全部为扩张卷积残差块，无池化层或其他层。在第 12 个扩张卷积残差块之后的 2 个全链接与原 Faster R-CNN 模型中 VGG16 后面的全连接层一致，神经元个数都为 4096，且在全连接层之后连接一个 dropout 层用以减轻网络模型对训练集的过拟合。由于本文行为识别的类别共 16 类，加上背景一共 17 类，因此“cls_score”全连接层种共 17 个神经元，“cls_prob”为一个 softmax 分类函数，输出目标属于 17 种类别的概率。“bbox_pred”全连接层中含有 68 个神经元，即目标对应于 17 种类别的目标框的回归信息。该融合网络中其他部分的连接方式与原 Faster R-CNN 网络模型中一致。

2.3 融合网络训练

本文的融合模型采用交替训练的方式对整个模型进行训练。

a) 单独训练 RPN 网络。该部分反向传播采用动量法(momentum)，本次 RPN 网络训练总的迭代次数为 30000 次，学习率衰减系数为 0.1，学习率衰减设置在第 20000 次迭代，即当迭代次数到 20 000 的时候，将学习率乘以衰减系数，得到新的学习率为 0.0001。采用原 Faster R-CNN 网络中训练好的 RPN 网络作为本文融合模型中 RPN 网络参数的初始值。

b)将第一步中训练 RPN 网络后输出的 256 个目标建议框带入融合网络的 DRN 网络部分，并单独训练该部分。该网络的损失函数见式(1)。由于融合模型 DRN 网络部分的复杂度远大于 RPN 网络部分，其模型达到收敛更加不易，因此总迭代次数设置为 60000 次，学习率衰减系数为 0.1，设置在第 50000 次迭代，momentum 系数为 0.9。采用均值和方差分别为 0 和 0.0001 的截断正态分布中的随机值作为初始值。

c)微调融合模型中的 RPN 网络部分，并且使融合网络中的 RPN 网络部分和 Fast R-CNN 网络部分共享卷积层，即图 5 中虚线框部分的第一个普通卷积层到第 12 个扩张卷积残差块。

d)微调融合模型中 Fast R-CNN 网络部分的全链接层，同样保持融合网络中的 RPN 网络部分和 Fast R-CNN 网络部分共享卷积层。

3 改进的融合模型

本章针对融合模型可能出现的“梯度消失”和“梯度爆炸”问题进行了两方面的改进。

3.1 添加 BN 层的融合模型

当网络的层数很深时，会出现“梯度消散”或“梯度爆炸”的现象。且会影响网络模型后面层的数据分布，在网络的训练中，若模型中的数据分布每次都不同，网络就需要不断的去拟合新的分布，导致网络的训练速度过慢。批量归一化(Batch Normalization,BN)可以有效的预防这个问题[26]。其原理即在深度网络模型的每一层之前添加一个可以学习的归一化层，表达式如下：

$$\begin{cases} \mu_\beta \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma_\beta^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}\left(x_i-\mu_\beta\right)^2 \\ \hat{x}_i \leftarrow \frac{x_i-\mu_\beta}{\sqrt{\sigma_\beta^2+\in}} \\ y_i \leftarrow \gamma \hat{x}_i+\beta \end{cases} \tag{5}$$

其中：x_i 为该批次(batch)的第 i 个输入，μ_β 为输入均值，σ_β^2 为方差，$\in$ 为一个很小的固定数，本文取 0.0001。$\hat{x}_i$ 为第 i 个新输入，y_i 为该 BN 层的输出，γ 和 β 都是该层中需要学习的参数。在测试阶段，BN 层的输出为

$$y=\frac{\gamma}{\sqrt{Var[x]+\in}}x+\left(\beta-\frac{\gamma E[x]}{\sqrt{Var[x]+\in}}\right) \tag{6}$$

其中：$E[x]$ 和 $Var[x]$ 分别为式(5)中所有批次的 μ_β 的均值和 σ_β 的无偏估计，其他参数含义与式(5)中一致。在图 5 的融合模型中，虚线框部分各层前面添加一个 BN 层，即全链接层和普通卷积层前添加一个 BN 层，在 DR1 至 DR12 的扩张卷积残差块中，两个卷积层的前面也添加一个 BN 层。对于 RPN 网络部分每个网络层前面同样添加 BN 层。

3.2 包含三层残差块的融合模型

原融合模型中扩张卷积残差块包含两层卷积层，对应图 4.1 中 DR1 至 DR12 部分。文献[9]提出了一种包含三层卷积层的扩张卷积残差块，结构如图 6 所示。图中的 x 为输入，relu 为斜坡函数。映射函数与两层的扩张卷积残差块一致，为 F(x)+x。本文采用的三个卷积层尺寸固定，第一个和第三个卷积层的尺寸为 1*1，其扩张值为 1。第二个卷积层的尺寸为 3*3，扩张值不固定。三个卷积层的滤波器个数不固定，步长为 1，填充全部采用“SAME”方式。三层扩张卷积残差块在网络层数很深的时候效果优于两层扩张卷积残差块。

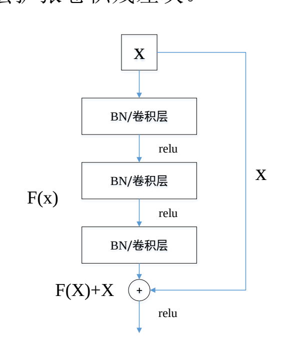


图 6 三层的扩张卷积残差块

将图 5 中虚线框部分 DR1 至 DR6 替换为三层扩张卷积残差块，“pool3”层后面的残差块保持不变，仍为两层扩张卷积残

差块。由于部分两层残差块替换为三层残差块，加深了网络模型，因此需要添加 BN 层，即网络各层前面增加一个 BN 层。

4 实验结果与分析

4.1 融合网络

本次实验在 GPU 版本的 tensorflow 上执行，该融合模型的训练共耗时约 3 天左右。得到训练好的模型之后，将部分示例图片带入到训练好的模型中测试其行为识别性能，示例图片通过该模型之后得到多个目标框及其在某一个类别上的得分，采用阈值为 0.7 的 NMS 剔除掉多余目标框后保留每类概率大于 0.8 的目标框。示例图片识别效果如下图。两张输入的示例图片尺寸不一致，图 7 尺寸为 600*450，两名篮球运动员被红色框标记出来，同时出了篮球行为的标签及概率值，即置信度。图 8 尺寸为 480*360，掷链球行为的概率大于阈值 0.8，同样被标记出来，然而该图中的其他人在 16 类行为上的概率未能超过 0.8，因此这些人物未被标记出来。

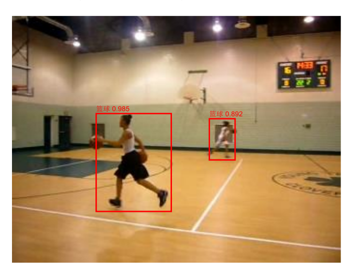


图 7 篮球行为识别示例



图 8 掷链球行为识别示例

检测完示例图片的识别效果之后，通过测试集计算该模型的 mAP，并与基于该数据库的其他行为识别算法比较，结果如表 1 所示。由表中可以看出，本文提出的融合模型在检测指标 mAP 上高于其他行为识别算法、原 Faster R-CNN 模型及采用本文所用数据库的 YOLO[26]和 SSD[27]算法。

表 1 本文算法及其他算法的 mAP

本文算法及相关算法	mAP
文献[21]	69.2%
文献[22]	76.4%
文献[23]	73.7%
文献[24]	72.3%
文献[25]	75.1%
文献[10]	72.1%
原 Faster R-CNN 模型	76.4%
YOLO	67.3%
SSD	76.8%
融合模型	77.2%

4.2 改进的融合模型与原融合模型对比

对两种改进的融合模型采用同样的方式进行训练，除改进部分外，其余的参数与原融合模型一致，在相同的测试集计算其 mAP 并进行对比。表 2 给出了实验对比结果。

表 2 两种改进的融合模型与原融合模型的 mAP

融合模型	mAP
融合模型	77.2%
添加 BN 层的融合模型	78.5%
含三层残差块的融合模型	78.9%

从表中可以看出，添加了 BN 层的融合模型的 mAP 较原融合模型有了一定的提升，达到了 78.5%，表明原融合模型中存在轻微的“梯度消失”或 “梯度爆炸”问题，而添加的 BN 层在一定程度上解决了该问题。包含三层扩张卷积残差块的融合模型识别效果最好，其 mAP 为 78.9%，表明本文所用的融合网络具有了一定的深度，此时三层的扩张卷积残差块在分类任务上的效果优于两层的扩张卷积残差块。

5 结束语

基于 DRN 网络在分类上的优势及 Faster R-CNN 网络在目标追踪上的精确性，本文将 DRN 网络部分的扩张卷积残差块引入到 Faster R-CNN 网络中代替原网络中的共享卷积层部分，形成一个融合网络。在该融合模型的基础上又提出了两种改进的融合模型：添加 BN 层的融合模型及包含三层扩张卷积残差块的融合模型。实验结果表明三种融合模型在分类指标 mAP 上均高于原 Faster R-CNN 模型及应用该数据库的其他行为识别算法，其中，包含三层扩张卷积残差块的融合模型取得了最高的 mAP，为 78.9%。但本文提出的融合模型在检测速度上略有欠缺，仅能达到每秒五帧左右，而 YOLO 和 SSD 算法均能达到每秒 45 帧及 58 帧左右。因此，如何在保证识别效果持续提升的同时，加快检测速度成为了今后的主要研究方向之一。

参考文献：

[1] 梅阳，王永雄，秦琪，等. 一种基于关键帧的人体行为识别方法 [J]. 光学技术, 2017, 43 (4): 323-328. (Mei Yang, Wang Yongxiong, Qing Qi, *et al*. A method of human behavior recognition based on key frame [J]. Optical Technology, 2017, 43 (4): 323-328.)

[2] 王忠民，张琮，衡霞，等. CNN 与决策树结合的新型人体行为识别方法研究 [J]. 计算机应用研究, 2017, 34 (12): 3569-3572. (Wang Zhongmin, Zhang Zong, Heng Xia, *et al*. Research on new human behavior recognition method combining CNN with decision tree [J]. Application Research of Computers, 2017, 34 (12): 3569-3572.)

[3] 肖玉玲. 结合 HOG//HOF 级联特征和多层分类器的人体行为识别 [J]. 计算机工程与设计, 2017, 38 (9): 2567-2572. (Xiao Yuling. Human behavior recognition combined with HOG//HOF cascade features and multilayer classifier [J]. Computer Engineering and Design, 2017, 38 (9):

2567-2572.)

[4] Quy N H, Quoc N H, Anh N T L, *et al.* 3D human face recognition using sift descriptors of face's feature regions [C]// Proc of the 1st IEEE International Conference on Computer Communication and the Internet. Cham: Springer, 2015: 117-126.

[5] Ayumi V, Fanany M I. A comparison of SVM and RVM for human action recognition [C]// Proc of Internationan Conference on Industrial Internet of Things. 2015.

[6] Prasad S, Ramkumar B. Passive copy-move forgery detection using SIFT, HOG and SURF features [C]// Proc of IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology. Cham: Springer, 2017: 706-710.

[7] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012: 1097-1105.

[8] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition [C]// Computer Vision and Pattern Recognition. 2016: 770-778.

[9] Yu Fisher, Koltun V, Funkhouser T. Dilated residual networks [J]. Computer Science, 2017: 636-644.

[10] Niebles J C, Chen C W, Li Feifei. Modeling temporal structure of decomposable motion segments for activity classification [C]// Proc of European Conference on Computer Vision. Springer-Verlag, 2010: 392-405.

[11] Koller D, Tang K, Li FeiFei. Learning latent temporal structure for complex event detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012: 1250-1257.

[12] Liu Jingen, Kuipers B, Savarese S. Recognizing human actions by attributes [C]// Computer Vision and Pattern Recognition. IEEE, 2011: 3337-3344.

[13] Liu Fang, Xu Xiangmin, Qing Chunmei. Temporal order information for complex action recognition [C]// Proc of IEEE International Conference on Consumer Electronics-China. IEEE, 2017.

[14] Razavian A S, Azizpour H, Sullivan J, *et al.* CNN features off-the-shelf: an astounding baseline for recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington DC: IEEE Computer Society, 2014: 512-519.

[15] Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2014: 580-587.

[16] Girshick R. Fast R-CNN [J]. Computer Science, 2015.

[17] Ren Shaoqing, He Kaiming, Girshick R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39 (6): 1137-1149.

[18] Zhang Yongbing, Sun Lulu, Wang Xingzhen. ReLU convolutional neutral network-based image denoising method: , CN 106204468 A [P]. 2016.

[19] 陈金辉，叶西宁. 行人检测中非极大值抑制算法的改进 [J]. 华东理工大学学报：自然科学版, 2015, 41 (3): 371-378. (Chen Jinghui, Ye Xining. Improvement of non-maximum value suppression algorithm in pedestrian detection [J]. Journal of East China University of Science and Technology: Natural Science Edition, 2015, 41 (3): 371-378.)

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014.

[21] Wang L, Yu Qiao, Tang X. Latent hierarchical model of temporal structure for complex activity classification. [J]. IEEE Trans on Image Processing, 2014, 23 (2): 810-22.

[22] Yan Shiyang, Smith J S, Lu Wenjin, *et al.* CHAM: action recognition using convolutional hierarchical attention model [J]. Computer Science. 2017.

[23] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention [J]. Computer Science, 2015.

[24] Liu Fang, Xu Xiangmin, Qiu Shuoyang, *et al.* Simple to complex transfer learning for action recognition. [J]. IEEE Trans on Image Processing, 2015, 25 (2): 949.

[25] Chen Xu, Hero A, Savarese S. Shrinkage optimized directed information using pictorial structures for action recognition [J]. Computer Science, 2014.

[26] Redmon J, Divvala S, Girshick R, *et al.* You only look once: unified, real-time object detection [C]// Computer Vision and Pattern Recognition. 2016: 779-788.

[27] Liu Wei, Anguelov D, Erhan D, *et al.* SSD: single shot multibox detector [J]. Computer Science. 2015: 21-37.

[28] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [J]. Computer Science. 2015: 448-456.